Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech

Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech

Yan Tang[a,b,*], Martin Cooke[c,a], Cassia Valentini-Botinhao[d]

[a] *Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain*
[b] *Acoustics Research Centre, University of Salford, UK*
[c] *Ikerbasque (Basque Science Foundation), Bilbao, Spain*
[d] *Centre for Speech Technology Research, University of Edinburgh, Edinburgh, UK*

Abstract

Several modification algorithms that alter natural or synthetic speech with the goal of improving intelligibility in noise have been proposed recently. A key requirement of many modification techniques is the ability to predict intelligibility, both offline during algorithm development, and online, in order to determine the optimal modification for the current noise context. While existing objective intelligibility metrics (OIMs) have good predictive power for unmodified natural speech in stationary and fluctuating noise, little is known about their effectiveness for other forms of speech. The current study evaluated how well seven OIMs predict listener responses in three large datasets of modified and synthetic speech which together represent 396 combinations of speech modification, masker type and signal-to-noise ratio. The chief finding is a clear reduction in predictive power for most OIMs when faced with modified and synthetic speech. Modifications introducing durational changes are particularly harmful to intelligibility predictors. OIMs that measure masked audibility tend to over-estimate intelligibility in the presence of fluctuating maskers relative to stationary maskers, while OIMs that estimate the distortion caused by the masker to a clean speech prototype exhibit the reverse pattern.

Keywords: Objective intelligibility metric, noise, speech modifications, synthetic speech

*Corresponding author at: Language and Speech Laboratory, Universidad del País Vasco, Vitoria, Spain.

Email address: y.tang@salford.ac.uk (Yan Tang)

1. Introduction

Spoken language applications using recorded natural[1] or synthetic speech can be made more robust through algorithmic speech modification. Unlike traditional speech enhancement techniques (e.g., Hu and Loizou, 2004; Martin, 2005; Chen et al., 2006; Srinivasan et al., 2007) which focus on the noise-corrupted speech signal, the speech modification approach (e.g., Sauert and Vary, 2006; Bonardo and Zovato, 2007; Yoo et al., 2007; Brouckxon et al., 2008; Tang and Cooke, 2010) alters the clean speech signal prior to output or transmission. A recent evaluation (Cooke et al., 2013b) demonstrated that speech modification can result in intelligibility gains in noise equivalent to increases of more than 5 dB in output level.

A key ingredient in the design of effective modification strategies is the estimation of listener performance at frequent intervals during the development cycle. However, while subjective intelligibility scores remain the ultimate reference, continuous behavioural testing during algorithm design is usually infeasible. An alternative is to use objective intelligibility metrics (OIMs) to predict listener scores. OIMs not only avoid the need for extensive subjective testing, but can also be used at the core of the algorithm optimisation process. A number of speech modification algorithms (e.g., Sauert and Vary, 2010a; Tang and Cooke, 2011; Taal et al., 2013; Valentini-Botinhao et al., 2014) have been developed and optimised based on maximising intelligibility predictions made by OIMs such as the Speech Intelligibility Index (SII; ANSI S3.5, 1997) or the glimpse proportion metric (GP; Cooke, 2006).

OIMs have been motivated by two distinct approaches to account for the effect of noise on speech. In addition to the aforementioned SII and GP metrics, the Articulation Index (AI; French and Steinberg, 1947; Fletcher and Galt, 1950; Kryter, 1962a,b), and the extended Speech Intelligibility Index (ESII; Rhebergen and Versfeld, 2005) focus on quantifying the *masked audibility* of speech in the presence of noise. On the other hand, techniques such as the Normalised-Covariance Measure (NCM; Holube and Kollmeier, 1996; Ma et al., 2009), the Christiansen-Pedersen-Dau metric (henceforth referred to as CPD for brevity; Christiansen et al., 2010) and the Short-Time Objective

[1]We use the term 'natural' to signify speech produced by a human talker as opposed to speech which is natural-sounding.

Intelligibility metric (STOI; Taal et al., 2010) correlate representations of the clean reference speech and the speech-plus-noise signal in an attempt to measure the *distortion* caused by the masker. Another distortion-based approach is the Coherence Speech Intelligibility Index (CSII) proposed by Kates and Arehart (2005). The CSII measures the similarity between clean and noisy speech using magnitude-square coherence (Carter et al., 1973; Kates, 1992) which quantifies the degree to which the output of a system is linearly related to its input.

Both audibility- and distortion-based approaches target spectro-temporal regions least affected by the noise, but differ in their assumptions. While techniques based on audibility require separated estimates of speech and noise in order to estimate masking, distortion-based OIMs assume that human listeners possess a template of the clean speech which is compared to the incoming noisy speech.

When an OIM is employed as the objective function to be maximised, the predictive accuracy of the OIM is critical in determining the validity and effectiveness of the optimisation process. Most of the OIMs mentioned above have been evaluated with recorded natural speech or speech processed by noise reduction techniques. Relatively few studies have investigated their predictive power for modified natural speech or synthetic speech in noise: most OIMs were originally proposed to predict the intelligibility of distorted natural speech, for distortions caused by additive noise together with artefacts introduced by suppression algorithms applied to the noisy speech signal.

Predicting the intelligibility impact of modification algorithms is likely to be challenging since the most successful methods (in terms of improving masked intelligibility) modify the signal in diverse domains – durational and spectral/formant – and possibly through non-linear operations. While the alterations benefit intelligibility, they may also introduce artefacts to the speech signal, leading to degraded speech quality. Nevertheless, the relation between speech intelligibility and quality is complex, and factors such as listening effort and loudness interact. Intelligibility and quality are not simply negatively or positively correlated, especially across listeners (Preminger and Tasell, 1995). For synthetic speech it might be expected that the OIMs' task is even more challenging because the natural speech reference signal is not available, i.e., distortions introduced by the Text-To-Speech (TTS) system cannot be taken into account. Consequently, predicting the intelligibility of poor quality synthetic speech may be even more difficult.

In two initial studies, which concerned solely the ability of OIMs to pre-

dict the masked intelligibility of modified and synthetic speech regardless of the perceptual speech quality, we observed a large reduction in the predictive accuracy of several OIMs on modified and synthetic speech relative to unmodified speech (Tang and Cooke, 2011; Valentini-Botinhao et al., 2011). The current study extends these pilots to a larger range of objective intelligibility metrics and includes behavioural data from recent extensive evaluations of 30 forms of modified and synthetic speech (Cooke et al., 2013a,b). Specifically, we evaluate the performance of one standard (SII) and six recent objective intelligibility metrics (ESII, GP, NCM, CSII, CPD, STOI) in predicting subjective intelligibility scores for both modified and synthetic speech in additive noise. The evaluation makes use of three datasets which together contain 396 combinations of speech modification, masker type and signal-to-noise ratio (SNR). The seven metrics are introduced in Section 2 while Section 3 describes the evaluation datasets. The outcome of a comparison of model predictions against behavioural data from large-scale listening tests is presented in Section 4.

2. Objective intelligibility models

2.1. Speech Intelligibility Index (SII)

SII and AI share a common underlying idea: speech intelligibility is dependent on the audibility of the signal in each frequency band. The AI can be expressed as a function of the masking level represented by the SNR (SNR_f^{AI}) in each frequency channel as

$$AI = \sum_{f=1}^{F} W_f \cdot SNR_f^{AI}, \quad \sum_{f=1}^{F} W_f = 1 \tag{1}$$

where W_f denotes the band importance function (BIF) in channel f and SNR_f^{AI} is a value in the interval $[0, 1]$ based on a piecewise-linear transformation of the actual SNR level SNR_f in band f

$$SNR_f^{AI} = \frac{\min(15, \max(-15, SNR_f)) + 15}{30} \tag{2}$$

SII extends AI by taking into account the effects of the upward spread of masking and high presentation levels when calculating the effective SNR in each band

$$SNR_f^{SII} = L_f \cdot \frac{\min(15, \max(-15, E_f - D_f)) + 15}{30} \tag{3}$$

where E_f and D_f denote the equivalent speech spectrum level and the disturbance spectrum level and L_f is a factor accounting for speech level distortion when speech is presented at high levels. The final SII is calculated as for the AI

$$SII = \sum_{f=1}^{F} W_f \cdot SNR_f^{SII} \tag{4}$$

In this study, SII is computed using 21 critical bands (i.e., $F = 21$) with the BIF for speech in noise from Tab. B.2 of ANSI S3.5 (1997).

2.2. Extended Speech Intelligibility Index (ESII)

ESII is an extension of SII designed to better predict intelligibility in the face of fluctuating maskers (Rhebergen and Versfeld, 2005). ESII computes the SII for each time frame (SII_{local}). Frame durations range from 35 ms for the lowest frequency critical band to 9.4 ms for the highest. The ESII model prediction is then based on the average SII across all frames:

$$ESII = \frac{\sum_{t=1}^{T} SII_{\text{local}}(t)}{T} \tag{5}$$

where T denotes the total number of frames. The procedure to calculate the local SII follows the original SII calculation described above.

2.3. Glimpse proportion (GP)

Simpson and Cooke (2005) compared the masking effectiveness of N-talker babble noise on speech intelligibility for a range of N, showing that a single competing speaker or amplitude-modulated noise is much less effective as a masker than multi-talker babble or speech-shaped noise. These basic findings motivated the glimpsing model of speech perception in noise (Cooke, 2006) in which not only temporal but also spectro-temporal modulations play a role in defining those parts of the speech signal that escape masking. The glimpse proportion is intended to reflect the local audibility of speech in noise and is defined as the percentage of spectro-temporal regions in modelled auditory excitation patterns whose local SNR exceeds a threshold α dB:

$$GP_{\text{original}} = \frac{100}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} \mathcal{H}(\, S_f(t) > (N_f(t) + \alpha)\,) \tag{6}$$

where T and F are the numbers of time frames and frequency channels, $S_f(t)$ and $N_f(t)$ denote the spectro-temporal excitation patterns (STEPs) of speech and noise at time t and frequency f, and $\mathcal{H}(.)$ is the Heaviside unit step function counting the number of 'glimpses' which meet the local masked audibility criterion α. The STEP is derived by a gammatone filterbank (Patterson et al., 1988) using an implementation introduced by Cooke (1993). The Hilbert envelope of each filter output is computed and smoothed by a leaky integrator with a 8 ms time constant (Moore et al., 1988), downsampled and log-compressed.

Glimpse proportion itself was not proposed originally as a dedicated intelligibility predictor but as an intermediate representation prior to a recognition component. However, with a number of simple extensions, GP has the potential to serve as an easily-computed proxy for the amount of speech that survives energetic masking. The new metric (Eq. 7) takes into account (i) the audibility of speech in quiet, (ii) the impact of durational changes, and (iii) ceiling performance, as detailed below:

$$GP = v[\frac{1}{T_{orig}F}\sum_{f=1}^{F}\sum_{t=1}^{T}\mathcal{H}(\,(S_f(t) > (N_f(t)+\alpha)) \wedge (S_f(t) > HL))] \qquad (7)$$

where $\wedge$ is logical conjunction and $v(.)$ is a quasi-log function defined as:

$$v(x) = \frac{\log(1+x/\delta)}{\log(1+1/\delta)},\ \delta = 0.01$$

To model audibility in quiet, $S_f(t)$ and $N_f(t)$ represent STEPs that have been adjusted by a frequency-dependent gain (ISO 389-7, 2006), a weighting that permits the use of a frequency-independent value of hearing level (HL), which is set to 25 dB here. To account for the decreased intelligibility of rapid speech, T_{orig} denotes the number of time frames in the unmodified speech STEP. The function v compresses GP scores to reflect the finding that subjective performance reaches a ceiling for GP values significantly lower than unity.

Here, a 34-channel gammatone filterbank with filter centre frequencies covering the range 100-7500 Hz linearly-spaced on the equivalent rectangle bandwidth scale (Moore and Glasberg, 1983) was used to derive the STEPs. The local masked audibility threshold α was set to 3 dB, a value which produced a high listener-model correlation ($\rho = 0.96$) in Cooke (2006).

2.4. Normalised-covariance measure (NCM)

Within the framework of the AI metric, the normalised-covariance measure (NCM) was motivated by the idea that noise both reduces and interferes with the temporal modulations of speech. Instead of measuring the SNR level, which is a relationship between speech and noise alone, the signal-to-distortion ratio (SDR) is used to quantify the degree of distortion. In frequency channel f the correlation coefficient r_f between the downsampled Hilbert envelopes of clean S_f and noisy speech Y_f is computed:

$$r_f = \frac{\sum_{t=1}^{T}(S_f(t) - \bar{S}_f) \cdot (Y_f(t) - \bar{Y}_f)}{\sqrt{\sum_{t=1}^{T}(S_f(t) - \bar{S}_f)^2 \cdot \sum_{t=1}^{T}(Y_f(t) - \bar{Y}_f)^2}} \tag{8}$$

where T denotes the length of the time series and $\bar{S}_f$ and $\bar{Y}_f$ are across-time averages of the clean and noisy speech envelopes in channel f. The SDR in decibel of channel f is defined as:

$$SDR_f^{\text{NCM}} = 10 \log_{10} \frac{r_f^2}{1 - r_f^2} \tag{9}$$

Following the SII, SDR_f^{NCM} is then transformed to a normalised index NI_f which lies in the range $[0, 1]$, using Eq. 2 with SNR_f^{AI} replaced by SDR_f^{NCM}. The final intelligibility index is computed using Eq. 1 as:

$$NCM = \sum_{f=1}^{F} W_f \cdot NI_f \tag{10}$$

where W_f denotes the BIF introduced along with the SII earlier. Ma et al. (2009) reported high correlations between listeners' sentence recognition scores and predicted intelligibility in babble noise ($\rho = 0.94$), car noise ($\rho = 0.85$), street noise ($\rho = 0.88$) and train noise ($\rho = 0.90$).

2.5. Coherence Speech Intelligibility Index (CSII)

CSII replaces the correlation coefficient between the frequency-dependent Hilbert envelopes of the clean and noisy speech in Eq. 9 with the magnitude-square coherence γ^2 to quantify the degree to which the noisy speech is linearly-related to the clean speech

$$|\gamma_k|^2 = \frac{\left|\sum_{t=1}^{T} S_k(t) Y_k^*(t)\right|^2}{\sum_{t=1}^{T} |S_k(t)|^2 \sum_{t=1}^{T} |Y_k(t)|^2} \tag{11}$$

where $S_k(t)$ and $Y_k(t)$ are FFT spectra in frame t of the speech and speech-plus-noise signals, k is the bin index, and T is the total number of frames. The quantity $|\gamma_k|^2$ is a value in the range $[0, 1]$. The SDR for a channel is defined as

$$SDR_f^{\text{CSII}} = 10 \log_{10} \frac{\sum_{k=1}^{K} R_f(k) \, |\gamma_k|^2 \, Y'(k)}{\sum_{k=1}^{K} R_f(k)[1 - |\gamma_k|^2] Y'(k)} \tag{12}$$

where f is the index of the simplified ro-ex filterbank R (Moore and Glasberg, 1983), K denotes the total number of FFT bins, and Y' is the noisy speech power spectral density, estimated using the FFT. The $CSII(t)$ of frame t is computed in each 16-ms Hamming window with a 50% overlap between windows using Eq. 1 and 2 with SNR_f^{AI} substituted by SDR_f^{CSII}.

To account for the differing degrees to which speech can be affected by the noise masker, frames are grouped into three levels – low, mid, high – according to the local root-mean-square energy (RMS in dB) relative to the overall RMS of the entire signal

$$RMS'(t) = 20 \log_{10} \frac{RMS(t)}{RMS_{overall}} \tag{13}$$

where $RMS(t)$ and $RMS_{overall}$ are the RMS of the amplitude of the signal waveform at frame t, and the entire signal, respectively. The low-level frames are those with a relative RMS range of -30 dB to -10 dB; -10 to 0 dB count as mid-level; those with positive relative RMS are classified as high-level frames. For each level, the mean $CSII_{high,mid,low}$ is obtained by averaging across all frames falling into this level:

$$\begin{cases} CSII_{high} = \frac{1}{T_{high}} \sum CSII(t) \text{ where } RMS'(t) \geqslant 0 \\ CSII_{mid} = \frac{1}{T_{mid}} \sum CSII(t) \text{ where } -10 \leqslant RMS'(t) < 0 \\ CSII_{low} = \frac{1}{T_{low}} \sum CSII(t) \text{ where } -30 \leqslant RMS'(t) < 10 \end{cases} \tag{14}$$

The final model output, $CSII$, is obtained by a linear weighting plus offset of the three level scores:

$$CSII = -3.47 + 1.84 \cdot CSII_{low} + 9.99 \cdot CSII_{mid} + 0.00 \cdot CSII_{high} \tag{15}$$

The weights shown minimise the mean-squared error between model predictions and listener scores, based on an unconstrained nonlinear minimisation method introduced by Nelder and Mead (1965). It can be seen that CSII

only uses the information from the mid and low-level frames to make intelligibility predictions. In Kates and Arehart (2005), CSII showed good intelligibility prediction ($\rho = 0.94$) for speech in additive noise and speech with peak-clipping and centre-clipping distortions.

2.6. The Christiansen-Pedersen-Dau metric (CPD)

CPD uses a psychoacoustically-validated model of auditory processing (Dau et al., 1996) to generate an internal representation of a signal. The signal is passed through 32 gammatone filters followed by half-wave rectification and five non-linear loops to model auditory nerve fibre adaptation. Denoting the internal representations of the reference speech signal and speech-noise mixture as s and y, a frame-based cross-correlation $r(t)$ between s and y is then computed every 20 ms with a 50% overlap:

$$r(t) = \frac{\sum_{f=1}^{F}\sum_{i=1}^{I}(\, y_f(i) - \bar{y}(t)\,)(\, s_f(i) - \bar{s}(t)\,)}{\sqrt{\sum_{f=1}^{F}\sum_{i=1}^{I}(\, y_f(i) - \bar{y}(t)\,)^2 \sum_{f=1}^{F}\sum_{i=1}^{I}(\, s_f(i) - \bar{s}(t)\,)^2}} \tag{16}$$

where $\bar{s}(t)$ and $\bar{y}(t)$ denote the average across time and frequency of the internal representation of the reference signal s and corrupted signal y at frame t, respectively. F and I are the number of frequency bands and samples in a frame. Simultaneously, the frames are classified into three levels (low, mid, high) following the frame classification introduced in CSII. The overall intelligibility of each level (r_{low}, r_{mid} and r_{high}) can be calculated by Eq. 14, except that the $CSII(t)$ of the frame t in Eq. 14 is substituted by the cross-correlation coefficients of that frame $r(t)$ computed by Eq. 16. Finally, the objective score CPD is obtained by a linear weighting of the three level scores:

$$CPD = w_{low}{\cdot}r_{low} + w_{mid}{\cdot}r_{mid} + w_{high}{\cdot}r_{high} \tag{17}$$

where w_{high}, w_{mid} and w_{low} are the weights for each level. In CPD the final intelligibility score actually only reflects the contribution of the high-level frames (i.e., $w_{low} = 0$, $w_{mid} = 0$ and $w_{high} = 1$). The intelligibility prediction by the CPD metric was reported in Christiansen et al. (2010) to have high correlation with subjective data in speech-shaped noise ($\rho = 0.96$), cafe noise ($\rho = 0.94$), car noise ($\rho = 0.97$) and bottle noise ($\rho = 0.88$) for enhanced speech processed by ideal time-frequency segregation (ITFS) techniques (Cooke et al., 2001; Wang, 2005).

2.7. Short-Time Objective Intelligibility (STOI)

The STOI metric was initially designed to predict the intelligibility of speech processed by enhancement algorithms such as ITFS, and a high correlation ($\rho = 0.95$) with subjective data was reported for this type of enhanced speech in noise (Taal et al., 2010). Both the corrupted speech signal and the reference signal are represented with time-frequency excitation spectra in 15 third-octave bands and 384-ms frames using the discrete Fourier transform. In order to de-emphasise the impact of regions in which noise dominates the spectrum, excitation spectra of the corrupted speech Y are further clipped by a normalisation procedure expressed in Eq. 18, where S is the excitation spectrum of the reference signal:

$$Y' = max(min(\lambda \cdot Y, (1 + 10^{-\beta/20}) \cdot S), (1 - 10^{-\beta/20}) \cdot S) \tag{18}$$

where

$$\lambda = \sqrt{\sum S_f(n)^2 / \sum Y_f(n)^2}$$

λ is a scale factor for normalising corrupted time-frequency bins in frequency band f and n is the time-frequency bin index. $\beta = -15$ dB denotes the lower SDR bound. Finally, intelligibility is predicted by the correlation coefficient between Y' and S averaged across all bands and frames:

$$STOI = \frac{1}{TF} \sum_{f=1}^{F} \sum_{t=1}^{T} r_f(t) \tag{19}$$

where T and F denote the total number of one-third octave bands and the total number of frames, and $r_f(t)$ is the local correlation coefficient between Y' and S at frequency f and time t.

3. Datasets

The OIMs described above were evaluated based on listeners' responses to speech from three datasets (Tab. 1). One – NATURAL– consists of unmodified and modified natural speech. A second dataset, TTS, contains speech generated by an HMM-based synthesiser. The third dataset, HURRICANE, is made up of both natural and synthetic speech. Further details of the listening tests are provided in the articles mentioned in Tab. 1.

Table 1: *Datasets used in the evaluation. NC and NL refer to the number of listening conditions and the number of listeners respectively in the cited behavioural studies.*

Dataset	Natural	Synthetic	NC	NL	Behavioural data
NATURAL	yes	no	24	24	Tang and Cooke (2011)
TTS	no	yes	192	88	Valentini-Botinhao et al. (2011)
HURRICANE	yes	yes	180	314	Cooke et al. (2013a,b)

3.1. NATURAL

The NATURAL dataset (Tang and Cooke, 2011) was created to investigate modifications to sentences drawn from the GRID corpus (Cooke et al., 2006) with a sampling frequency of 25 kHz. Twenty-four native British English speakers identified the spoken letter ('A–Z', except 'W') and digit (0–9) keywords in sentences such as 'Place blue at C 3 now'. Listeners listened to the sentences embedded in noise over headphones, and chose letter and digit options from an onscreen keyboard. The percentage of keywords recognised correctly was taken as the measure of intelligibility.

Six speech conditions were compared: unmodified speech + five modification strategies. In Tang and Cooke (2011) a sixth modification approach was also evaluated, but led to low subjective scores for the SNRs employed, and is not included in the current evaluation. Two techniques equalised SNR in each frame or frequency channel; a third approach transferred energy to time-frequency bins just below the threshold of audibility; another introduced a pause placed to avoid epochs of intense noise, while the final technique combined the latter two approaches. The unmodified and modified utterances were mixed with speech-shaped noise (SSN) and speech-modulated noise (SMN) at two global SNRs of -9 and -6 dB, leading to a total of 24 listening conditions (unmodified + 5 modifications $\times$ 2 masker types $\times$ 2 SNR levels). The SSN noise has the long-term spectrum of the corpus; the SMN noise is generated from the SSN by modulating with the envelope of a randomly-concatenated utterance from the same corpus.

3.2. TTS

The TTS dataset, described in Valentini-Botinhao et al. (2011), is based on the responses of 88 listeners to text-to-speech (TTS) utterances presented

to listeners in various noise conditions over headphones. After each sentence listeners typed the words they heard in the sentence. The subjective intelligibility in each condition was computed as the percentage of correctly identified words.

TTS samples with a sampling frequency of 20 kHz were generated using a HMM-based speech synthesis system (Zen et al., 2009), which was trained with 4000 sentences uttered by a male British English speaker in the 'rjs' corpus from the University of Edinburgh. The synthetic voice has a similar quality to that generated by the HTS2005 system in the Blizzard Challenge 2010 (King and Karaiskos, 2010), with a mean opinion score (MOS) of 2.5. Synthetic speech was further processed using four different approaches to simulate the acoustic properties of natural Lombard speech; each process had three parameter settings. These processes and accompanying settings are: spectral peak enhancement (no enhancement and two enhancement levels), fundamental frequency change (one decreasing and two increasing), shifts in the line spectral pairs domain (three different shifts amounts, always towards the high frequency region) and speech rate changes (one faster and two slower). Processed synthetic speech was then added to four maskers at four different SNRs chosen to produce word accuracies of approximately 20, 40, 60, and 80%: speech-shaped noise (SSN: −11.8, −8.8, −6.2 and −3.1 dB), cafeteria babble noise (BAB: −9.5, −6.8, −4.6 and −1.9 dB), car noise (CAR: −31.9, −28.4, −25.5 and −22.0 dB) and high frequency noise (HiFQ: −43.5, −37.6, −32.7 and −26.8 dB). Consequently, the TTS dataset contains speech material for 192 listening conditions (4 processes × 3 parameter settings × 4 maskers × 4 SNRs).

3.3. HURRICANE

The third dataset is based on the 2012 and 2013 Hurricane Challenges, the results of which are presented in Cooke et al. (2013b) and Cooke et al. (2013a), respectively. The subjective data was from 314 native British English speakers listening to the Harvard sentences (Rothauser et al., 1969) uttered by a male British English speaker over headphones. Each sentence contains five to six keywords such as 'the juice of lemons makes fine punch'; as above, listener performance was assessed using the keyword identification rate.

The speech type in this dataset consists of both natural (plain and Lombard) speech, modified speech and synthetic speech, at a sampling frequency of 16 kHz. The synthetic speech has a MOS of 3.0 in terms of speech quality

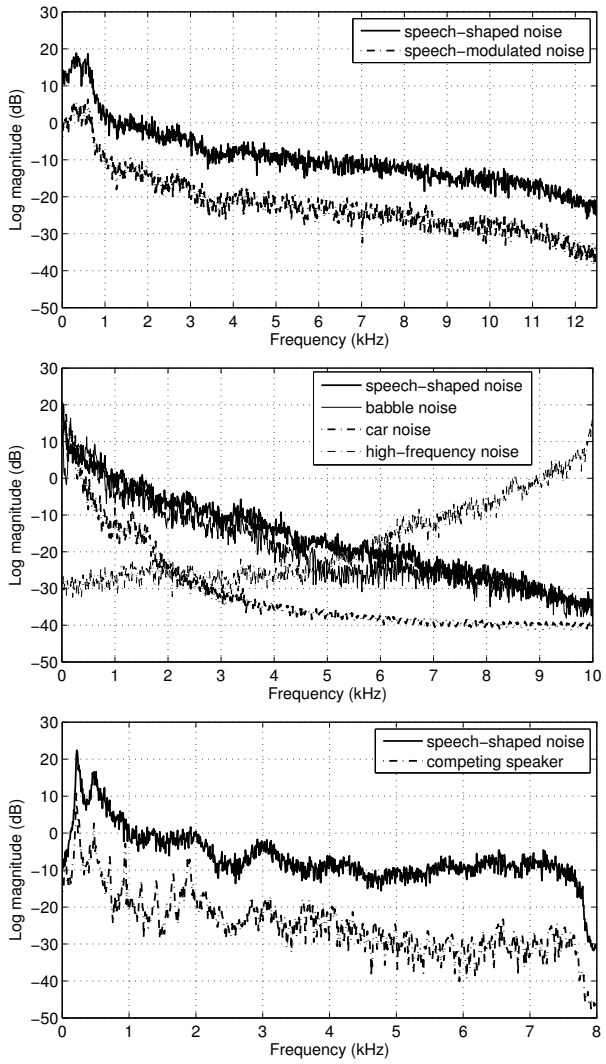

Figure 1: *LTAS of the noise maskers used in the* NATURAL *(top),* TTS *(middle) and* HURRICANE *(bottom) datasets. For clarity, LTAS of speech-modulated noise in the* NATURAL *dataset and the competing speech masker in the* HURRICANE *dataset have been offset by -10 dB.*

(Chen et al., 2014). In all, 30 types of speech were evaluated in two maskers: competing speech (CS) spoken by a female speaker, and SSN, both at three SNR levels chosen to produce recognition scores of about 25, 50 and 75% in each noise masker; specific SNRs are −21, −14, −7 dB for CS and −9, −4, 1 dB for SSN. The HURRICANE dataset thus represents 180 listening

conditions (30 modifications × 2 maskers × 3 SNRs). Appendix A provides a brief summary of the 30 modification techniques.

Long term average spectra (LTAS) of the maskers used in the three datasets are shown in Fig. 1. In the three experiments, the subjective intelligibility of each test condition was calculated as the mean word recognition rate across all listeners.

4. Objective intelligibility predictions

All OIMs were evaluated by inspecting both the Pearson correlation coefficient ρ between mean listener scores and the raw output of the metric, and the standard deviation of the error σ_e, computed as

$$\sigma_e = \sigma_d \cdot \sqrt{1 - \rho^2} \tag{20}$$

where σ_d is the standard deviation of subjective intelligibility scores for a given experimental condition. Statistical comparisons among dependent correlations were conducted using a method described in Meng et al. (1992) based on Chi-squared tests on z-transformed scores.

4.1. Overall performance of each metric

Table 2 reports correlations ρ and the standard deviations of the error σ_e across all modifications and noise maskers for each of the three datasets. Since distortion-based OIMs require a reference signal – normally clean speech – as input to the metric, a choice must be made between using the clean unmodified or clean modified speech as the reference. For modified speech whose duration is altered by modification algorithm, the modified clean speech itself always has to be used as reference because the distortion-based OIMs require the reference signal to have the same duration as the tested signal. Outcomes using clean unmodified and clean modified speech as the reference signal are shown in the table. Since most OIMs make better predictions using modified clean speech as a reference, this is used in subsequent comparisons.

For the NATURAL and TTS datasets, most of the chosen models performed significantly less well with modified natural speech and TTS speech than reported in previous studies for unmodified natural speech or noisy speech processed by noise reduction algorithms. The performance of the models varied significantly in the NATURAL case $[\chi^2(6) = 64.895, p < 0.001]$: while GP, CPD, SII and ESII performed similarly $[Z = 0.731, p = 0.465]$,

Table 2: *Correlations (ρ) and standard deviation of the error (σ_e) between listeners' scores and OIM predictions for the three datasets. OIMs above the divider are audibility-based while those below are distortion-based. For the latter group, results are shown for both unmodified and modified clean speech reference signals.*

		NATURAL	TTS	HURRICANE
		ρ (σ_e)	ρ (σ_e)	ρ (σ_e)
SII		0.77 (0.11)	0.78 (0.15)	0.68 (0.19)
ESII		0.82 (0.09)	0.67 (0.18)	0.67 (0.20)
GP		0.89 (0.07)	0.78 (0.15)	0.66 (0.20)
NCM:	(mod)	-0.10 (0.16)	0.79 (0.15)	0.51 (0.23)
	(unmod)	0.05 (0.17)	0.73 (0.17)	0.50 (0.23)
CSII:	(mod)	0.56 (0.14)	0.60 (0.19)	0.76 (0.17)
	(unmod)	0.36 (0.15)	0.51 (0.21)	0.75 (0.17)
CPD:	(mod)	0.79 (0.10)	0.73 (0.17)	0.83 (0.15)
	(unmod)	0.31 (0.14)	0.65 (0.18)	0.83 (0.15)
STOI:	(mod)	0.15 (0.16)	0.62 (0.19)	0.63 (0.20)
	(unmod)	0.12 (0.16)	0.63 (0.19)	0.62 (0.21)

audibility-based OIMs (i.e., GP, SII and ESII) outperformed those based on distortion (i.e., NCM, CSII and STOI) [$Z = 6.568, p < 0.001$], with the exception of CPD. For the TTS dataset, the models also performed differently [$\chi^2(6) = 57.573, p < 0.001$]: the quality of predictions made by NCM, ESII, GP and CPD were statistically-equivalent [$Z = 0.778, p = 0.437$] and superior to those of the remaining OIMs [$Z = -9.109, p < 0.001$]. The performance of the metrics also differed for the HURRICANE dataset [$\chi^2(6) = 76.155, p < 0.001$]. Here, CSII and CPD were equivalent [$Z = 1.777, p = 0.076$] and superior to the other metrics [$Z = 6.422, p < 0.001$].

Figures 2, 3 and 4 present individual data points for each condition in the three datasets, colour-coded by masker type. The CPD metric shows the most uniform performance across the datasets, but even this model lacks predictive power, especially for the TTS dataset. For the HURRICANE dataset, all OIMs apart from CPD fare badly when predicting the intelligibility for both maskers combined: while ESII and GP predict higher-than-actual intelligibility for the competing speech masker, SII, NCM, STOI and to a lesser extent CSII show the converse pattern.

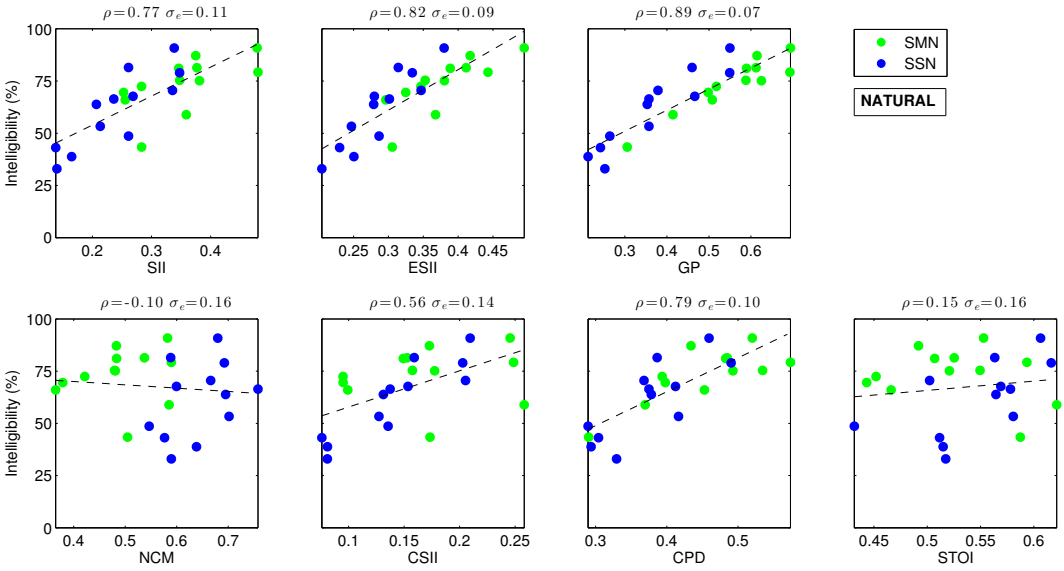

Figure 2: *Subjective intelligibility scores versus OIM predictions for the* NATURAL *dataset.*

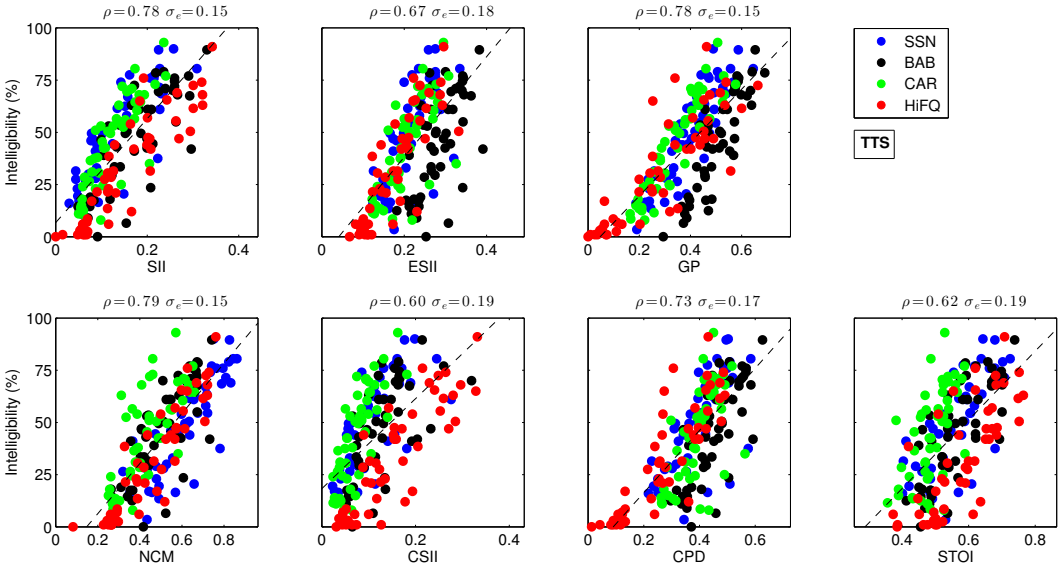

Figure 3: *Subjective intelligibility scores versus OIM predictions for the* TTS *dataset.*

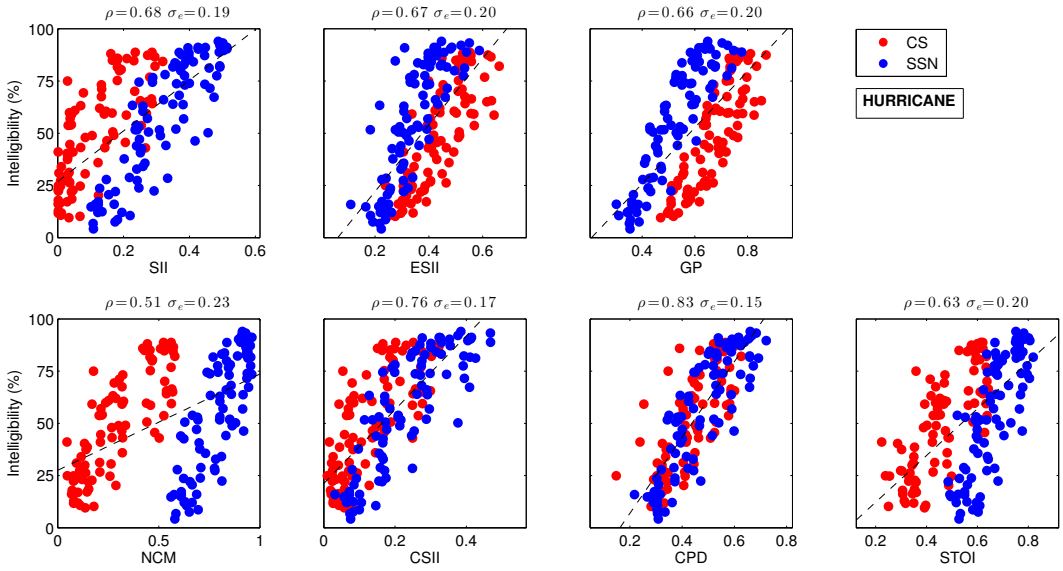

Figure 4: *Subjective intelligibility scores versus OIM predictions for the* HURRICANE *dataset.*

4.2. Performance for sub-conditions

4.2.1. NATURAL

Fig. 5 shows the breakdown by noise type for the NATURAL dataset, and reveals that the GP metric is most correlated with listeners ($\rho > 0.90$) for both maskers, followed by ESII and CPD. Many of the metrics make reasonable ($\rho > 0.80$) intelligibility predictions for speech in the presence of stationary maskers. Apart from GP, predictions were less good for the modulated masker than for stationary maskers, especially for NCM, CSII and STOI. Except for CPD, distortion-based OIMs produced lower correlations for the modulated maskers [$Z = -6.701, p < 0.001$]. The poor performance of these three OIMs for this dataset may be due to them being unable to deal with the modification which introduced pauses into the speech signal.

4.2.2. TTS

Correlations split by masker type in the TTS dataset are shown in Fig. 6. As expected, correlations are generally higher for individual maskers than overall, suggesting a high degree of masker-specificity in the ability of metrics to predict intelligibility. Most metrics made reasonable predictions ($\rho > 0.85$) for the high frequency noise masker. For the remaining maskers the pat-

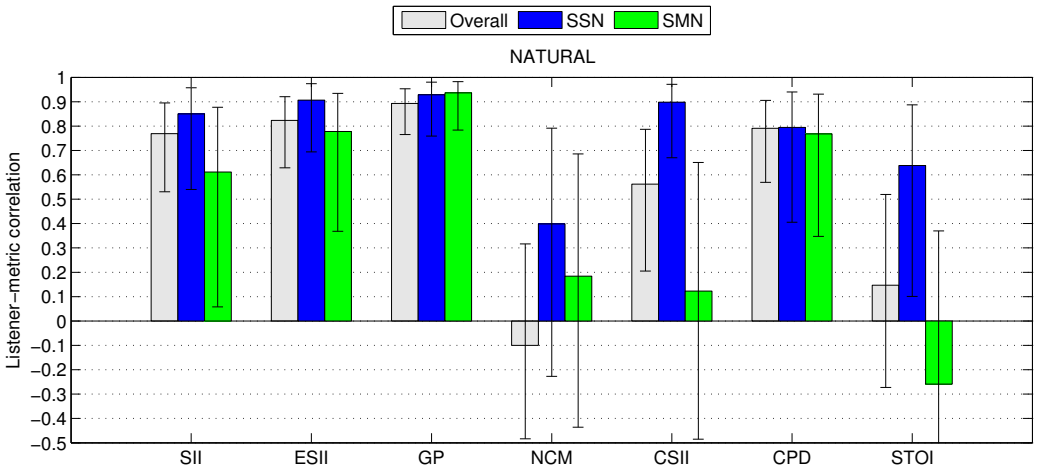

Figure 5: *Correlations for the* NATURAL *dataset, both overall and split by subsets of individual noise maskers. Here and elsewhere error bars indicate 95% confidence intervals.*

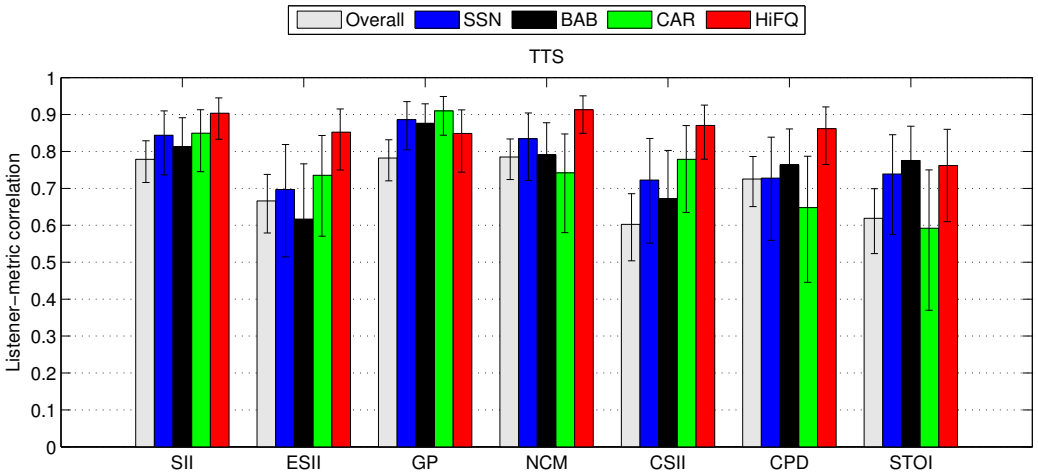

Figure 6: *Correlations for the* TTS *dataset, both overall and split by subsets of individual noise maskers.*

tern of correlation showed significant variation across metrics [for all maskers $\chi^2(6) \geq 25.062, p < 0.001$].

Tab. 3 provides a more detailed look at correlations for the different processes applied during synthesis, relative to a synthetic speech baseline. The performance of most OIMs decreased as a result of processing. Apart from SII, all OIMs struggled to predict the effect of `LSP shift`. Other than

Table 3: *Correlations for unprocessed and processed synthetic speech from the* TTS *dataset.*

	unprocessed ρ (σ_e)	**peak** ρ (σ_e)	**F0 change** ρ (σ_e)	**LSP shift** ρ (σ_e)	**temporal change** ρ (σ_e)
SII	0.85 (0.11)	0.85 (0.12)	0.81 (0.13)	0.86 (0.13)	0.61 (0.20)
ESII	0.80 (0.13)	0.78 (0.14)	0.78 (0.14)	0.75 (0.17)	0.37 (0.24)
GP	0.84 (0.12)	0.81 (0.13)	0.82 (0.13)	0.76 (0.17)	0.83 (0.14)
NCM	0.93 (0.08)	0.90 (0.10)	0.89 (0.10)	0.76 (0.17)	0.76 (0.17)
CSII	0.62 (0.17)	0.65 (0.17)	0.68 (0.16)	0.54 (0.21)	0.49 (0.22)
CPD	0.89 (0.10)	0.85 (0.12)	0.86 (0.11)	0.78 (0.16)	0.47 (0.22)
STOI	0.81 (0.13)	0.70 (0.16)	0.73 (0.15)	0.70 (0.18)	0.46 (0.23)

GP, the performance of all OIMs degraded drastically in the face of linear expansion and contraction of duration.

4.2.3. HURRICANE

All metrics produced higher correlations in stationary noise than in the fluctuating competing speaker condition in the HURRICANE dataset (Fig. 7). OIMs were further evaluated for three groupings of conditions: natural speech only, synthetic speech only and for conditions that changed speech duration (Tab. 4). Except for CSII in for synthetic speech [$Z = 1.009, p = 0.313$] and modifications resulting in temporal change [$Z = 0.908, p = 0.364$], CPD demonstrated more robust predictive power [for remaining conditions $Z \geq 2.477, p < 0.05$]. It is perhaps surprising to see that, in general, the subset of synthetic speech is better predicted than the natural speech subset. One possible explanation is that the degree of processing applied in the TTS dataset was more extreme than seen in the HURRICANE dataset. Modifications involving temporal changes also appear less harmful in the HURRICANE dataset than in TTS particularly for those metrics – ESII, CSII, CPD and STOI – that suffered most. Again, this is likely to be due to the scale of temporal changes involved in the two datasets. In the HURRICANE dataset, speech duration was expanded by a smaller factor than in the TTS data, and in no condition was duration decreased.

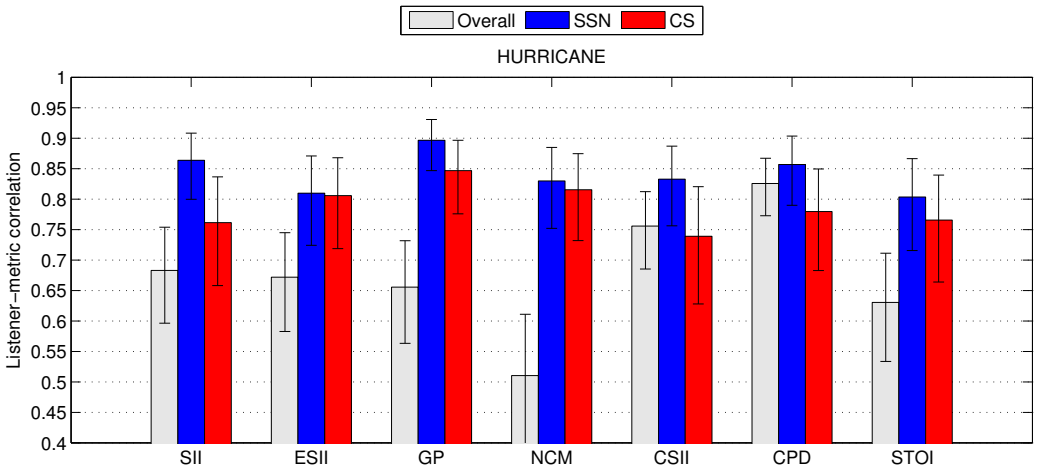

Figure 7: *Correlations for the* HURRICANE *dataset, both overall and split by subsets of individual noise maskers.*

Table 4: *Listener-model correlations for natural speech, synthetic speech and speech whose duration is altered by the modifications from the* HURRICANE *dataset (3.3).*

	natural ρ (σ_e)	**synthetic** ρ (σ_e)	**temporal change** ρ (σ_e)
SII	0.70 (0.19)	0.74 (0.15)	0.68 (0.18)
ESII	0.69 (0.19)	0.79 (0.14)	0.66 (0.19)
GP	0.68 (0.19)	0.73 (0.15)	0.68 (0.18)
NCM	0.54 (0.22)	0.54 (0.19)	0.50 (0.22)
CSII	0.77 (0.17)	0.89 (0.10)	0.75 (0.16)
CPD	0.87 (0.13)	0.94 (0.08)	0.81 (0.15)
STOI	0.67 (0.19)	0.66 (0.17)	0.63 (0.19)

5. Discussion

Compared to model-listener correlations reported in the literature for unmodified natural speech or speech processed by noise reduction techniques, the current study highlights a clear reduction in the performance of a representative range of OIMs for modified and synthetic speech. One contributing factor for most OIMs is their inability to predicting intelligibility across different maskers, especially for stationary versus highly-fluctuating maskers. Additionally, many OIMs were adversely affected by modifications involv-

ing temporal changes such as pause insertion and duration alteration. The following sections explore these two issues further.

5.1. Intelligibility predictions across maskers

To be considered robust, an intelligibility metric ought to cope with different masking conditions. In the current study, short-term audibility-based metrics (i.e., GP and ESII) tended to over-predict intelligibility in fluctuating maskers and under-predict in stationary noise, while the converse was the case for SII and the distortion-based metrics. Both GP and ESII use short-term information to quantify the degree to which spectro-temporal regions of the speech dominate those from the masker in terms of energy. An interesting question is whether all such target-dominant regions contribute equally to intelligibility in the face of different maskers. Fluctuating noise provides more opportunities for glimpsing spectro-temporal regions of the target signal due to masker envelope modulations than stationary masker. However, neither GP nor ESII take the effects of non-simultaneous (e.g., forward masking) and informational masking into account, and as a consequence may overestimate intelligibility for fluctuating noise maskers.

Since speech is more tolerant of energetic masking in fluctuating noise than in stationary noise at the same SNR level, it is necessary to present speech in a fluctuating masker at a lower SNR to obtain the same intelligibility level as that of speech presented in a stationary masker with the same long-term spectrum. Given that SII predictions are based on the long-term spectral SNR, SII is sensitive to any change in global SNR, which may explain why SII scores are lower in the CS than in the SSN condition. The same considerations apply to NCM and CSII which inherit the SII framework. Further, NCM, CSII and STOI compute correlation across time only and hence do not account for across-frequency distortion. Based on the findings here, it appears that measuring distortion solely in the time domain exaggerates the negative impact of a fluctuating noise masker on intelligibility.

CPD was the only distortion-based metric which predicted intelligibility across masker types reasonably well. This may be due to the use of cross-correlation in both time and frequency domains, as well as the adoption of a sophisticated auditory representation. To explore these hypotheses further, we isolated the components in the auditory model (Dau et al., 1996) which generates the internal representation used in CPD. Fig. 8 presents a performance breakdown using the output of the auditory model with different component combinations.

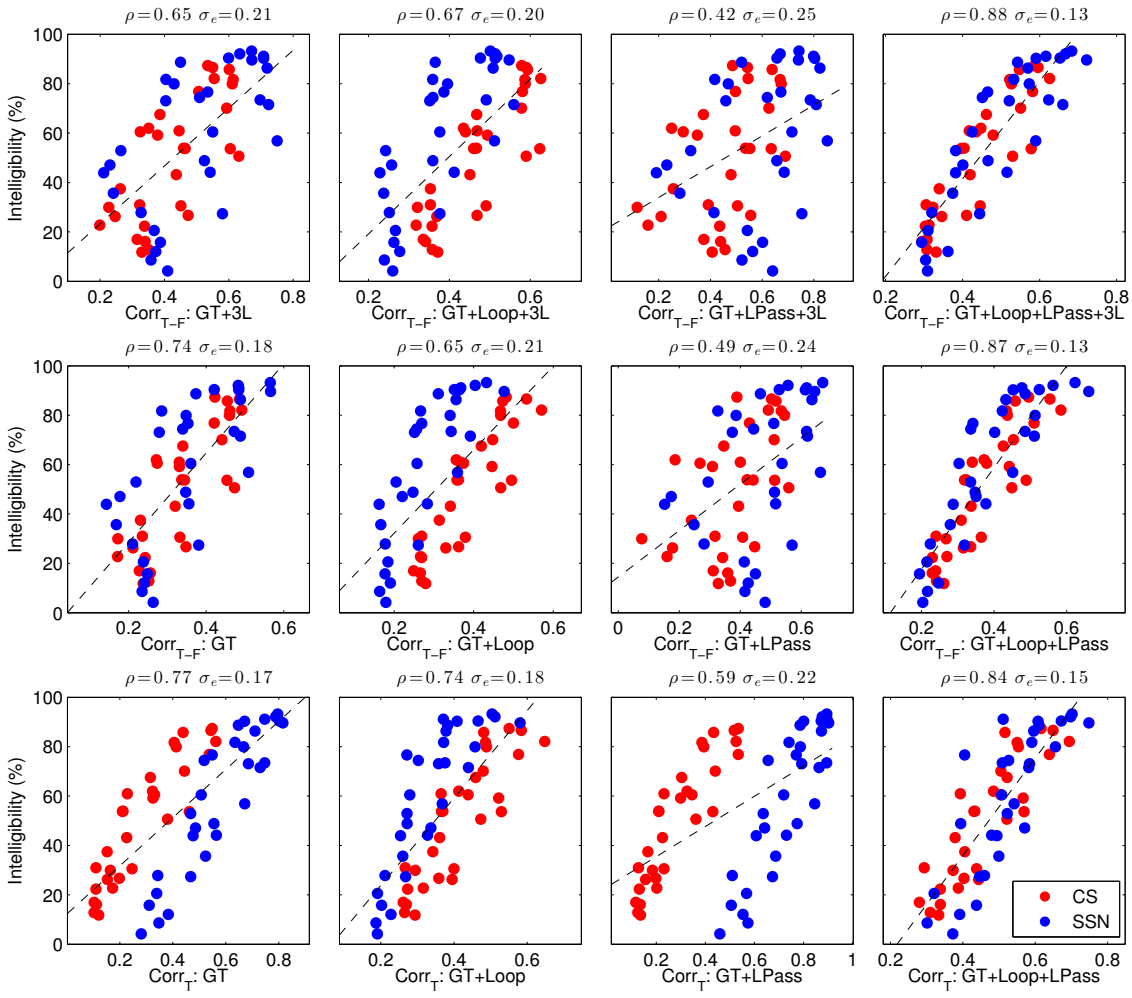

Figure 8: *Performance of CPD with different component combinations.* ***Corr$_{T-F}$****: correlation across time and frequency;* ***Corr$_T$****: correlation across time frames;* ***GT****: output of gammatone filters;* ***Loop****: non-linear adaptation loops;* ***LPass****: low-pass filter extracting the envelope;* ***3L****: 3-level criteria. The top row illustrates the performance of CPD with different components when the correlation is computed across time and frequency; the middle row shows results without applying 3-level criteria; the bottom row shows the effect of computing correlation across time only, without applying the 3-level criteria.*

After gammatone filtering (`GT` in Fig. 8), the auditory model has two major components: the non-linear adaptation loops which model non-simultaneous masking (`Loop`), followed by the low-pass filter which extracts their envelope (`LPass`); the CPD metric finally makes the prediction with 3-level criteria

(3L; Eq. 14) as shown in the top row of Fig. 8. If intelligibility predictions are made using only the output of the gammatone filterbank, the method of computing correlations (Eq. 16) seems to matter, since cross-correlating time and frequency leads to similar predictions for different maskers (column 1, middle row), whereas correlating across only time in each frequency independently (column 1, lower row) results in under-estimation for CS and over-estimation for SSN as seen in other distortion-based OIMs such as STOI and NCM. The synergy of the non-linear adaptation loops (column 2) and low-pass filter (column 3) is strong; neither of the two components alone is able to achieve a similar performance to that of their combination (column 4). The additional impact of the 3-level criterion (column 4, upper row) is rather small, perhaps because only a limited number of noise conditions were used to demonstrate the effects here in comparison to the evaluation conducted by Christiansen et al. (2010). Nevertheless, the use of 3-level criteria is not harmful.

5.2. Effects of durational change

Few OIMs were capable of predicting the intelligibility of speech processed by temporal modifications in the TTS dataset. An assumption is implied in the GP metric (Eq. 7) that listeners benefit from a slow speech rate, presumably because it allows listeners longer to process the received information, or increases the likelihood of glimpsing, particularly in fluctuating maskers. Applying a single temporal change factor to model this assumption helps improve the predictive accuracy of the GP metric in this study. It is natural to ask whether the performance of other OIMs also be improved by weighting their output by a temporal change factor d (Eq. 21):

$$d = \frac{D_{mod}}{D_{orig}} \tag{21}$$

where D_{mod} and D_{orig} are the durations of the modified and original speech signals. Tab. 5 shows correlations before and after applying the durational factor. It is clear that all the OIMs evaluated here do indeed benefit – substantially in most cases – from this simple assumption.

There is some evidence that clear and slow speech is beneficial to speech intelligibility in noise (Picheny et al., 1985; Payton et al., 1994; Uchanski et al., 1996). However, simple linear or non-linear time-scale modifications on the speech signal seem not to be able to improve speech in noise performance compared to non-processed speech (Schmitt, 1983; Nejime and Moore,

Table 5: *Listener-model correlations for modifications from the* TTS *dataset that altered duration, with and without the temporal change factor d.*

	Without ρ (σ_e)	**With** ρ (σ_e)
SII	0.61 (0.20)	0.80 (0.15)
ESII	0.37 (0.24)	0.82 (0.15)
GP	0.61 (0.20)	0.83 (0.14)
NCM	0.76 (0.17)	0.80 (0.15)
CSII	0.49 (0.22)	0.70 (0.18)
CPD	0.47 (0.22)	0.82 (0.14)
STOI	0.46 (0.23)	0.69 (0.18)

1998). It has been argued that this is because simple changes in speaking rate on their own may not be able to provide novel contributions on which to base intelligibility improvements; these may come instead from other acoustic changes in clear speech compared with normal speech (Smiljanić and Bradlow, 2009), such as vowel space expansion (Picheny et al., 1986), increased F0 (Bradlow et al., 2003) and a change in spectral tilt (Krause and Braida, 2004) for example.

Valentini-Botinhao et al. (2011) tested the intelligibility of synthetic speech at three speech rates: 0.6, 1.4 and 2.0 (calculated as the ratio presented in Eq. 21) relative to a normal speech rate of 1.0. A rate = 0.6 indicates a faster speech rate, whereas rate = 1.4 and rate = 2.0 lead to slower speech. As expected, increasing speech rate harmed the intelligibility of synthetic speech. This could be because some harmful acoustic changes (e.g., a decrease in the quality of transitions between syllables) were inevitably induced when the signals were synthesised according to the demands of speech rate. The results presented in Valentini-Botinhao et al. (2011) also suggest that decreasing speech rate may to some extent help listeners understand synthetic speech better in noise. When the speech rate decreased too much, the benefit to speech intelligibility ceased. We are not aware of any studies which suggest the point at which a decrease in speech rate starts to compromise the positive contributions to intelligibility from other factors. One possibility is that a listener's perception is negatively affected by changes in coarticulation and the artefacts that might have been introduced by over-long durations, as

well as the unfamiliar speaking style. It is thus not clear whether applying the simple durational correction factor of Eq. 21 will always improve OIM performance.

5.3. Speech quality

Several studies have shown a correlation between subjective and objective measures for quality and intelligibility prediction. Early studies concerned quality prediction for speech coders (Barnwell III, 1980; Quackenbush et al., 1988; Kubichek et al., 1991). More recently-introduced measures and evaluation methods are designed to measure the quality and intelligibility of noise-corrupted speech processed using noise reduction algorithms (Hu and Loizou, 2006, 2008; Ma et al., 2009; Taal et al., 2009; Ma and Loizou, 2011; Gomez et al., 2012) and dereverberation algorithms (Kokkinakis and Loizou, 2011). An issue for the future is to determine how well metrics are able to account for any changes to the quality of both modified and synthetic speech in noise.

6. Conclusions

In the current study state-of-the-art OIMs that provide good predictions of natural speech performed less well for modified and synthetic speech, especially for those modifications introducing temporal changes. While many OIMs produced reasonable estimates for modified speech in the presence of single masker types, across-noise predictions were generally poor. Methods motivated by masked audibility tended to over-estimate intelligibility for fluctuating maskers and under-estimate intelligibility for stationary maskers, while for many metrics that computed the distorting effect of noise on clean speech the reverse pattern was evident. These findings suggest that further development of OIMs is required to enable their use in applications such as the offline development and online deployment of speech modification algorithms.

Acknowledgements

This study was supported by the LISTA Project (http://listening-talker.org), funded by the Future and Emerging Technologies programme within the 7th

Framework Programme for Research of the European Commission, FET-Open grant number 256230. We thank Yannis Stylianou for sharing a MATLAB implementation of ESII, and Cees Taal for making the MATLAB implementation of STOI available online for free access. The implementation of SII is available online at `http://www.sii.to` while MATLAB implementations of NCM and CSII are provided in the CD accompanying the book (Loizou, 2013). The MATLAB implementation of the GP metric can be obtained by request to the first author.

Appendix A. Speech conditions in the HURRICANE dataset

Table A.6: *Techniques marked with an asterisk are reported in Cooke et al. (2013b); the remaining algorithms are described in Cooke et al. (2013a). A subscript of 'd' indicates a modification that alters the duration of the speech signal. Note that there are two 'Plain' and two 'TTS' conditions.*

Natural	
Plain*	Unmodified natural speech
Plain	Unmodified natural speech
Lombard$^*_d$	Lombard speech
SelBoost*	Boosting just-audible time-frequency regions (Tang and Cooke, 2010)
OptGP*	Glimpse-optimised spectral reweighting (Tang and Cooke, 2012)
OptSII*	SII-optimised spectral reweighting (Sauert and Vary, 2010a,b)
TMDRC*	Harmonic model tilt modification + dynamic range compression (DRC) (Erro et al., 2012)
SSDRC*	Spectral shaping + DRC(Zorila et al., 2012)
F_0-shift	Optimised energetic masking by shifting F0 (Villegas and Cooke, 2012)
GCRetime$_d$	Modify local speech rate + preserve information (Aubanel and Cooke, 2013)
phoneLLabso$_d$	ASR based phone energy equalisation (Zhang et al., 2013)
phoneLLdscr$_d$	ASR based contextual phone energy equalisation (Petkov and Kleijn, 2013)
uwSSDRCt$_d$	SSDRC + time-scaling, vowel space expansion and transients enhancement (Godoy and Stylianou, 2013)
AdaptDRC	SII-based adaptive DRC (Schepker et al., 2013)
IWFEMD	Modification of intrinsic mode function of empirical mode decomposition
on/offset	Temporal energy reallocation to consonant bursts and vocalic onsets
OptimalSII	SII-optimised time-invariant spectral reallocation (Taal and Jensen, 2013)
RESSTSMOD	Loudness increase based on source-filter modelling
SBM	Spectral binary masking

SEO	Spectral energy optimisation by emphasising perceptually motivated acoustic features (Takou et al., 2013)
SINCoFETS$_d$	Local time-scaling + DRC + psychoacoustic adaptive equalisation (Brouckxon and Verhelst, 2013)
SSS	Temporal energy reallocation based on steady-state suppression (Hodoshima et al., 2006)
Synthetic	
TTS$^*_d$	HMM-based text-to-speech
TTS$_d$	HMM-based text-to-speech
TTSLomb$^*_d$	TTS adapted to Lombard (Valentini-Botinhao et al., 2012b)
TTSGP$^*_d$	Glimpse-optimised TTS (Valentini-Botinhao et al., 2012a,c)
PSSDRC-syn	HMM synthesis + noise-independent modifications at vocoder level (Erro et al., 2013)
TTSLGP-DRC$_d$	Lombard adaptation, spectral envelope optimisation + DRC (Valentini-Botinhao et al., 2013)
C2H-TTS$_d$	Phonetic contrast maximisation (Nicolao et al., 2012)
GlottLombard$_d$	Lombard adaptation using glottal inverse filtering + DRC and formant sharpening (Suni et al., 2013)

ANSI S3.5, 1997. Methods for the calculation of the Speech Intelligibility Index.

Aubanel, V., Cooke, M., 2013. Information-preserving temporal reallocation of speech in the presence of fluctuating maskers. In: Proc. Interspeech. pp. 3592–3596.

Barnwell III, T., 1980. Correlation analysis of subjective and objective measures for speech quality. In: Proc. ICASSP. pp. 706–709.

Bonardo, D., Zovato, E., 2007. Speech synthesis enhancement in noisy environments. In: Proc. Interspeech. pp. 2853–2856.

Bradlow, A. R., Kraus, N., Hayes, E., 2003. Speaking clearly for learning-impaired children: Sentence perception in noise. J. Speech Hear. Res. 46, 80–97.

Brouckxon, H., Verhelst, W., 2013. An overview of the VUB entry for the 2012 Hurricane Challenge. In: Proc. Interspeech. pp. 3602–3604.

Brouckxon, H., Verhelst, W., Schuymer, B. D., 2008. Time and frequency dependent amplification for speech intelligibility enhancement in noisy environments. In: Proc. Interspeech. pp. 557–560.

Carter, G. C., Knapp, C. H., Nuttall, A. H., 1973. Estimation of the magnitude-squared coherence function via overlapped fast Fourier transform processing. IEEE Trans. Audio Electroacoust. 21, 337–344.

Chen, J., Benesty, J., Huang, Y., Doclo, S., 2006. New insights into the noise reduction Wiener filter. IEEE Trans. Audio, Speech, and Language Processing 14 (1), 1218–1234.

Chen, L.-H., Raitio, T., Valentini-Botinhao, C., Yamagishi, J., Ling, Z.-H., September 2014. DNN-Based Stochastic Postfilter for HMM-Based Speech Synthesis. In: Proc. Interspeech. Singapore, pp. 1954–1958.

Christiansen, C., Pedersen, M. S., Dau, T., 2010. Prediction of speech intelligibility based on an auditory preprocessing model. Speech Comm. 52 (7-8), 678–692.

Cooke, M., 1993. Modelling Auditory Processing and Organisation. Cambridge University Press.

Cooke, M., 2006. A glimpsing model of speech perception in noise. J. Acoust. Soc. Am. 119 (3), 1562–1573.

Cooke, M., Barker, J., Cunningham, S., Shao, X., 2006. An audio-visual corpus for speech perception and automatic speech recognition. J. Acoust. Soc. Am. 120 (5), 2421–2424.

Cooke, M., Mayo, C., Valentini-Botinhao, C., 2013a. Intelligibility-enhancing speech modifications: the Hurricane Challenge. In: Proc. Interspeech. pp. 3552–3556.

Cooke, M., Mayo, C., Valentini-Botinhao, C., Stylianou, Y., Sauert, B., Tang, Y., 2013b. Evaluating the intelligibility benefit of speech modifications in known noise conditions. Speech Comm. 55, 572–585.

Cooke, M. P., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Comm. 34 (3), 267–285.

Dau, T., Puschel, D., Kohlrausch, A., 1996. A quantitative model of the "effective" signal processing in the auditory system. I. Model structure. J. Acoust. Soc. Am. 99 (6), 3615–3622.

Erro, D., Stylianou, Y., Navas, E., Hernaez, I., 2012. Implementation of Simple Spectral Techniques to Enhance the Intelligibility of Speech using a Harmonic Model. In: Proc. Interspeech. pp. 639–642.

Erro, D., Zorila, T.-C., Stylianou, Y., Navas, E., Hernaez, I., 2013. Statistical synthesizer with embedded prosodic and spectral modifications to generate highly intelligible speech in noise. In: Proc. Interspeech. pp. 3557–3561.

Fletcher, H., Galt, R. H., 1950. The perception of speech and its relation to telephony. J. Acoust. Soc. Am. 22, 89–151.

French, N. R., Steinberg, J. C., 1947. Factors governing the intelligibility of speech sounds. J. Acoust. Soc. Am. 19 (1), 90–119.

Godoy, E., Stylianou, Y., 2013. Increasing speech intelligibility via spectral shaping with frequency warping and dynamic range compression plus transient enhancement. In: Proc. Interspeech. pp. 3572–3576.

Gomez, A., Schwerin, B., Paliwal, K., 2012. Improving objective intelligibility prediction by combining correlation and coherence based methods with a measure based on the negative distortion ratio. Speech Comm. 54 (3), 503–515.

Hodoshima, N., Arai, T., Kusumoto, A., Kinoshita, K., 2006. Improving syllable identification by a preprocessing method reducing overlap-masking in reverberant environments. J. Acoust. Soc. Am. 119, 4055–4064.

Holube, I., Kollmeier, B., 1996. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. J. Acoust. Soc. Am. 100, 1703–1716.

Hu, Y., Loizou, P., 2008. Evaluation of objective quality measures for speech enhancement. IEEE Trans. Audio, Speech, and Language Processing 16 (1), 229–238.

Hu, Y., Loizou, P. C., 2004. Speech enhancement based on wavelet thresholding the multitaper spectrum. IEEE Trans. Speech Audio Processing, 59–67.

Hu, Y., Loizou, P. C., 2006. Evaluation of objective measures for speech enhancement. In: Proc. Interspeech. pp. 1447–1450.

ISO 389-7, 2006. Acoustics – Reference Zero For The Calibration Of Audiometric Equipment – Part 7: Reference Threshold Of Hearing Under Free-field And Diffuse-field Listening Conditions.

Kates, J., Arehart, K., 2005. Coherence and the speech intelligibility index. J. Acoust. Soc. Am. 117 (4), 2224–2237.

Kates, J. M., 1992. On using coherence to measure distortion in hearing aids. J. Acoust. Soc. Am. 91, 2236–2244.

King, S., Karaiskos, V., Sept. 2010. The Blizzard Challenge 2010. Kyoto, Japan.

Kokkinakis, K., Loizou, P. C., 2011. Evaluation of objective measures for quality assessment of reverberant speech. In: Proc. ICASSP. pp. 2420–2423.

Krause, J. C., Braida, L. D., 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. J. Acoust. Soc. Am. 115, 362–378.

Kryter, K. D., 1962a. Methods for the calculation and use of the Articulation Index. J. Acoust. Soc. Am. 34, 1689–1697.

Kryter, K. D., 1962b. Validation of the articulation index. J. Acoust. Soc. Am. 34, 1698–1702.

Kubichek, R., Atkinson, D., Webster, A., 1991. Advances in objective voice quality assessment. In: Glob. Telecomm. Conf. Vol. 3. pp. 1765–1770.

Loizou, P. C., 2013. Speech Enhancement: Theory and Practice, Second Edition. CRC Press.

Ma, J., Hu, Y., Loizou, P. C., 2009. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. J. Acoust. Soc. Am. 125 (5), 3387–3405.

Ma, J., Loizou, P. C., 2011. SNR loss: A new objective measure for predicting the intelligibility of noise-suppressed speech. Speech Comm. 53 (3), 340–354.

Martin, R., 2005. Speech Enhancement Based on Minimum Mean-Square Error Estimation and Supergaussian Priors. IEEE Trans. Speech Audio Processing 13 (5), 845–856.

Meng, X., Rosenthal, R., Rubin, D. B., 1992. Comparing correlated correlation coefficients. Psychological Bulletin 111 (1), 172–175.

Moore, B. C. J., Glasberg, B. R., 1983. Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. J. Acoust. Soc. Am. 74, 750–753.

Moore, B. C. J., Glasberg, B. R., Plack, C. J., Biswas, A. K., 1988. The shape of the ear's temporal window. J. Acoust. Soc. Am. 83 (7-8), 1102–1116.

Nejime, Y., Moore, B. C. J., 1998. Evaluation of the effect of speech-rate slowing on speech intelligibility in noise using a simulation of cochlear hearing loss. J. Acoust. Soc. Am. 103, 572–576.

Nelder, J. A., Mead, R., 1965. A simplex method for function minimization. Comput. J. 7, 308–313.

Nicolao, M., Latorre, J., Moore, R. K., 2012. C2H: A computational model of H&H-based phonetic contrast in synthetic speech. In: Proc. Interspeech. Portland, USA.

Patterson, R. D., Holdsworth, J., Nimmo-Smith, I., Rice, P., 1988. SVOS Final Report: The Auditory Filterbank. Technical report 2341, MRC Applied Psychology Unit.

Payton, K. L., Uchanski, R. M., Braida, L. D., 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. J. Acoust. Soc. Am. 95, 1581–1592.

Petkov, P. N., Kleijn, W. B., 2013. Preservation of speech spectral dynamics enhances intelligibility. In: Proc. Interspeech. pp. 3597–3601.

Picheny, M. A., Durlach, N. I., Braida, L. D., 1985. Speaking clearly for the hard of hearing. I. Intelligibility differences between clear and conversational speech. J. Speech Hear. Res. 28, 96–103.

Picheny, M. A., Durlach, N. I., Braida, L. D., 1986. Speaking clearly for the hard of hearing. II. Acoustic characteristics of clear and conversational speech. J. Speech Hear. Res. 29, 434–446.

Preminger, J. E., Tasell, D. J. V., 1995. Quantifying the relation between speech quality and speech intelligibility. J. Speech Hear. Res. 38 (3), 714–725.

Quackenbush, S., Barnwell, T., Clements, M., 1988. Objective measures of speech quality. Prentice Hall, New Jersey, USA.

Rhebergen, K. S., Versfeld, N. J., 2005. A Speech Intelligibility Index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners. J. Acoust. Soc. Am. 117 (4), 2181–2192.

Rothauser, E. H., Chapman, W. D., Guttman, N., Silbiger, H. R., Hecker, M. H. L., Urbanek, G. E., Nordby, K. S., Weinstock, M., 1969. IEEE Recommended practice for speech quality measurements. IEEE Trans. Audio Electroacoust 17, 225–246.

Sauert, B., Vary, P., 2006. Near end listening enhancement: speech intelligibility improvement in noise environments. In: Proc. ICASSP. pp. 493–496.

Sauert, B., Vary, P., 2010a. Near end listening enhancement optimized with respect to Speech Intelligibility Index and audio power limitations. In: Proc. EUSIPCO. Aalborg, Denmark, pp. 1919–1923.

Sauert, B., Vary, P., 2010b. Recursive closed-form optimization of spectral audio power allocation for near end listening enhancement. In: Proc. ITG-Fachtagung Sprachkommunikation. Bochum, Germany.

Schepker, H., Rennies, J., Doclo, S., 2013. Improving speech intelligibility in noise by sii-dependent preprocessing using frequency-dependent amplification and dynamic range compression. In: Proc. Interspeech. pp. 3577–3581.

Schmitt, J. F., 1983. The effects of time compression and time expansion on passage comprehension by elderly listeners. J. Speech Lang. Hear. Res. 26, 373–377.

Simpson, S. A., Cooke, M., 2005. Consonant identification in N-talker babble is a nonmonotonic function of N. J. Acoust. Soc. Am. 118 (5), 2775–2778.

Smiljanić, R., Bradlow, A. R., 2009. Speaking and Hearing Clearly: Talker and Listener Factors in Speaking Style Changes. Language and Linguistics Compass 3 (1), 236–264.

Srinivasan, S., Samuelsson, J., Kleijn, W., 2007. Codebook-Based Bayesian Speech Enhancement for Nonstationary Environments. IEEE Trans. Audio, Speech, and Language Processing, 441–452.

Suni, A., Karhila, R., Raitio, T., Kurimo, M., Vainio, M., Alku, P., 2013. Lombard modified text-to-speech synthesis for improved intelligibility: Submission for the Hurricane Challenge 2013. In: Proc. Interspeech. pp. 3562–3566.

Taal, C., Hendriks, R., Heusdens, R., Jensen, J., Kjems, U., 2009. An evaluation of objective quality measures for speech intelligibility prediction. In: Proc. Interspeech. pp. 1947–1950.

Taal, C., Jensen, J., 2013. SII-based speech preprocessing for intelligibility improvement in noise. In: Proc. Interspeech. pp. 3582–3586.

Taal, C. H., Hendriks, R. C., Heusdens, R., Jensen, J., 2010. A short time objective intelligibility measure for time-frequency weighted noisy speech. In: Proc. ICASSP. pp. 4214–4217.

Taal, C. H., Jensen, J., Leijon, A., 2013. On Optimal Linear Filtering of Speech for Near-End Listening Enhancement. IEEE Trans. Signal Process. Lett. 20 (3), 225–228.

Takou, R., Seiyama, N., Imai, A., 2013. Improvement of speech intelligibility by optimization of spectral energy. In: Proc. Interspeech. pp. 3605–3607.

Tang, Y., Cooke, M., 2010. Energy reallocation strategies for speech enhancement in known noise conditions. In: Proc. Interspeech. pp. 1636–1639.

Tang, Y., Cooke, M., 2011. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In: Proc. Interspeech. pp. 345–348.

Tang, Y., Cooke, M., 2012. Optimised spectral weightings for noise-dependent speech intelligibility enhancement. In: Proc. Interspeech. pp. 955–958.

Uchanski, R. M., Choi, S., Braida, L. D., Reed, C. M., Durlach, N. I., 1996. Speaking clearly for the hard of hearing. IV. Further studies of the role of speaking rate. J. Speech Hear. Res. 39, 494–509.

Valentini-Botinhao, C., Maia, R., Yamagishi, J., King, S., Zen, H., 2012a. Cepstral analysis based on the Glimpse proportion measure for improving the intelligibility of HMM-based synthetic speech in noise. In: Proc. ICASSP. pp. 3997–4000.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2011. Can objective measures predict the intelligibility of modified HMM-based synthetic speech in noise? In: Proc. Interspeech. Florence, Italy, pp. 1837–1840.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012b. Evaluating speech intelligibility enhancement for HMM-based synthetic speech in noise. In: Proc. SAPA Workshop.

Valentini-Botinhao, C., Yamagishi, J., King, S., 2012c. Mel cepstral coefficient modification based on the Glimpse Proportion measure for improving the intelligibility of HMM-generated synthetic speech in noise. In: Proc. Interspeech. pp. 631–634.

Valentini-Botinhao, C., Yamagishi, J., King, S., Maia, R., 2014. Intelligibility enhancement of HMM-generated speech in additive noise by modifying Mel cepstral coefficients to increase the glimpse proportion. Computer Speech & Language 28 (2), 665–686.

Valentini-Botinhao, C., Yamagishi, J., King, S., Stylianou, Y., 2013. Combining perceptually-motivated spectral shaping with loudness and duration modification for intelligibility enhancement of HMM-based synthetic speech in noise. In: Proc. Interspeech. pp. 3567–3571.

Villegas, J., Cooke, M., 2012. Maximising objective speech intelligibility by local f0 modulation. In: Proc. Interspeech. pp. 1704–1707.

Wang, D. L., 2005. Speech Separation by Humans and Machines. Springer US, Ch. On ideal binary mask as the computational goal of auditory scene analysis, pp. 181–197.

Yoo, S. D., Boston, J. R., El-Jaroudi, A., Li, C., Durrant, J. D., Kovacyk, K., S., S., 2007. Speech signal modification to increase intelligibility in noisy environments. J. Acoust. Soc. Am. 122 (2), 1138–1149.

Zen, H., Tokuda, K., Black, A., 2009. Statistical parametric speech synthesis. Speech Comm. 51 (11), 1039–1064.

Zhang, M., Petkov, P., Kleijn, B., 2013. Rephrasing-based speech intelligibility enhancement. In: Proc. Interspeech. pp. 3587–3591.

Zorila, T., Kandia, V., Stylianou, Y., 2012. Speech-in-noise intelligibility improvement based on spectral shaping and dynamic range compression. In: Proc. Interspeech. Portland, USA.